EUROGRAPHICS 2006 / E. Gröller and L. Szirmay-Kalos
(Guest Editors)

Volume 25 (2006), Number 3

Projective Alpha Colour

Philip Willis

Media Technology Research Centre,
Department of Computer Science, University of Bath, UK
http://www.bath.ac.uk/media/

Abstract

Alpha colours were introduced for image compositing, using a pixel coverage model. Algebraically they resemble homogeneous coordinates, widely used in projective geometry calculations. We show why this is the case. This allows us to extend alpha beyond compositing, to all colour calculations regardless of whether pixels are involved and without the need for a coverage model. Our approach includes multi-channel spectral calculations and removes the need for 7 channel and 6 channel alpha colour operations. It provides a unified explanation of pre-multiplied and non pre-multiplied colours, including negative coordinates and infinite points in colour space. It permits filter and illumination operations. It unifies the three existing significant compositing models in a single framework. It achieves this with a physically-plausible energy basis.

Keywords: projective geometry, homogeneous coordinates, image compositing, alpha blending, alpha compositing, colour representation, filtering, spectral colour, projective alpha colour

1. Introduction

It is usual to present alpha colours in the pre-multiplied form:

$$(\alpha r, \alpha g, \alpha b, \alpha) \qquad \alpha \neq 0$$

The standard compositing formulae use the colour components only when multiplied by their alpha value, so it is sensible to store and manipulate colours that way. In the homogeneous representation of coordinates for projective geometry, a similar form appears:

$$(wx, wy, wz, w) \qquad w \neq 0$$

In this case the w factor is effectively the local scale of the projective space. Computer graphics exploits this to produce 2D perspective drawings from objects defined in 3D but embedded in this 4D space. Again there are advantages, centred around the use of 4D matrix transformations.

This similarity of form has been apparent since 1984, when Porter and Duff's alpha compositing paper [PD84] appeared. In 1994, Jim Blinn [Bli94] wrote about alpha:

> This looks suspiciously like homogeneous coordinates. I've tried real hard, but for the life of me I can't figure out any use for this observation.

Blinn's implied challenge has remained an open problem since, one which we now solve as one aspect of a wider approach. Our formulation offers the following *key points*.

1. We demonstrate that alpha colours form a projective space, valid for any colour computation and not just those for compositing pixels.
2. We provide a unified explanation of pre- and non pre-multiplied colours and of colour normalisation.
3. We interpret colours in $[0.0, 1.0]$ and colours in $[-\infty, \infty]$ in a single model.
4. We remove the need for 5×5 matrices for colour translation operations.
5. We explain $\alpha, R, G, B > 1.0$.
6. We extend alpha to multi-channel spectral work, with a single alpha value.
7. We eliminate the need for 7 channel (*RGBARGB*) and 6 channel (*RGBAAA*) models.
8. We give a practical explanation of $\alpha, R, G, B < 0.0$.
9. We unify the compositing work of Wallace [Wal81], of Porter and Duff [PD84] and of Oddy and Willis [OW91], in a single framework.

To achieve this, we develop a homogeneous coordinate projective space for colour, using an energy density model.

2. The Projective Space

Projective 4-space can be thought of as Euclidean 3-space extended with the points and lines at infinity [Her92]. The finite points are called "affine" and the infinite points are called "ideal". This 4-space has certain properties which we will need to use. These are:

Property 1 Two distinct points define a line.

Property 2 Two distinct coplanar lines intersect at a point.

Property 3 In homogeneous coordinates, the set of projective points $(wx, wy, wz, w), w \neq 0$ are equivalent to the single Euclidean point (x, y, z).

These properties, with suitable variation, apply regardless of the dimensionality of the projective space. Property 1 applies whether the points are affine or ideal, in any combination. Property 2 might produce an affine point or an ideal point. This ability to cope with infinite points is of direct relevance in computer graphics, where parallel lines in 3D might project onto the image as intersecting lines. The intersection is called the *vanishing point* and can be thought of as the finite projected image of the infinite point where the parallel lines "meet". For the projective plane there is only one such point: it does not matter in which direction we follow the parallel lines. More formally, this is the ideal point in the direction of the parallel lines.

The natural transformations of projective spaces are those that turn lines into lines. From this it can be shown that they turn intersection points into intersection points. We also note that when homogeneous coordinates are used, these transformations (and only these) can be represented by matrices. Properties 1 and 2 tell us that we are working with a projective space.

Property 3 is especially interesting to us. We do not need homogeneous coordinates to discuss projective geometry but they are an attractive way to do so. In our case they turn out to be fundamental to alpha compositing and this property is a key one. It says that, given a Euclidean point (x, y, z) the line of points (wx, wy, wz, w) are all the same projective point, even as w varies. If $w = 0$ however, we can no longer tell which of the lines passing through the origin it falls on; in fact it falls on them all. We therefore have to be careful about w, which has a role distinct from (x, y, z).

We now argue that alpha colours form a projective space. We can clearly make a lexical substitution $(\alpha r, \alpha g, \alpha b, \alpha)$ but there also has to be a basis in the laws of physics. (We will not consider aspects of colour which relate to the human eye or to human psychology; rather, we are concerned with practical issues of colour manipulation in computing hardware and software.) We therefore need to give physical meaning to (r, g, b) and to α.

In colour, we often use (r, g, b) coordinates. These can be thought of as a special case of multi-channel spectral coordinates, with the dimensionality set to three. As above, we will from time to time appeal to (r, g, b) for its familiarity but our working will not assume it. Initially we will show a colour C, meaning a vector. Each component of this vector will be a measure of *energy* in a selected frequency band. This is our physical interpretation of the colour dimensions.

We expect this (r, g, b) space to be Euclidean. We note that it offers ideas of points, lines, planes and volumes; and that these all have a ready physical interpretation in colour. Pairs of distinct coplanar lines might intersect at a point but they do not do so if the lines are parallel. Parallel lines are colour lines separated by a constant colour. Distinct planes intersect at a line, unless they too are parallel. Importantly, C is unbounded; that is, we are not considering the unit colour cube, as we might for a pixel image or a display, but one in which indefinitely large finite values may occur. There seems to be a problem with interpreting negative colours but we argue that these are simply colours which reduce the energy component of a sum of colours. (We will give a stronger justification and a more practical example later.) This is the analogue of negative distances reducing a sum of distances. We do indeed have a Euclidean colour space.

We next turn to alpha. The w dimension is effectively a dimension of scale, with larger w normalising to give smaller Euclidean results. The case of $w = 0$ cannot be used this way but, because it identifies ideal points, it never has to be used to normalise. We interpret alpha physically as the general *volume* containing the colour energy. If we are interested in volume data, it will be the 3D volume and so on for higher dimensions. If we are interested in surfaces in 3D or in 2D images, alpha amounts to *area*. Note however we are not defining it as the area of a pixel or as pixel coverage but as general area. Negative areas might seem to be a problem but we will treat these as something which reduces a sum of areas. (We will again give a stronger justification later.)

We now investigate whether (C, α) can be interpreted as a projective space. First we address infinite points, which are the added feature from the Euclidean case. Parallel Euclidean colour lines have a clear interpretation, being two coplanar lines separated by a constant colour: both lines share the same direction. In what sense can we say that this direction defines an infinite colour point? The interpretation is simply that, as both lines "reach" infinite energy, the finite energy difference between them is insignificant; both are the same colour at the infinite energy point. Moreover that colour is well defined, by the direction, and is distinct from that of any other direction.

Now we can ask if the three properties given earlier can be interpreted with these (C, α) colours?

Property 1 Two distinct points define a line.
Any two finite (affine) colours can be joined by a line. The line represents the linear interpolation of the colours, both within and beyond. If one point is ideal and one affine then we have a point and a direction, which defines a line. If both

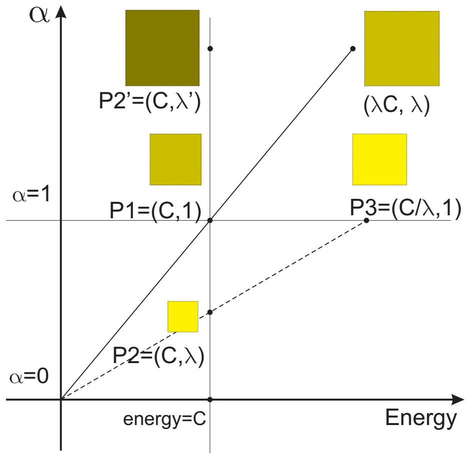

Figure 1: *Projective colour space*

are ideal, then the line is the line at infinity. Each of these results has a clear physical interpretation.

Property 2 Two distinct coplanar lines intersect at a point. This too has a clear interpretation at finite points. The infinite point case is justified by the direction argument just given, except when one of the lines is the line at infinity. In that case, the direction of the first line fixes an ideal point which, also being on the line at infinity, is the intersection point.

Property 3 In homogeneous coordinates, the projective points $(wx, wy, wz, w), w \neq 0$ are equivalent to the Euclidean point (x, y, z). If we scale both the colour and the alpha, then we are keeping fixed the *energy density*. So the points $(\alpha C, \alpha)$ are equivalent to $(C, 1)$; that is, they are all the same colour (C) in Euclidean space.

In our interpretation, an alpha colour records the energy of the colour and the area (volume) it is held in. The energy density is energy per unit area, which is directly proportional to intensity and can be thought of as such for our purposes. It follows that we can scale the homogeneous coordinates of a colour and still have the same intensity (Property 3).

We conclude that alpha colours (C, α) have an underlying projective space (*key point 1* in the Introduction) and that we can meaningfully interpret them in homogeneous coordinates. In turn this means we can use matrix operations. We will now explore the practical consequences of this.

3. Projective Alpha Colours, PACs

In Figure 1, if we place the original energy (C) at $P1 = (C, 1)$, then we are positioning this colour energy in a unit area. If instead we place it anywhere on the line $(\lambda C, \lambda)$, then we are changing both the energy and the area in proportion. We can think of each as falling on a central projection and generated from:

$$(C, 1)\begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix} = (\lambda C, \lambda)$$

All these points have the same intensity. They are called *pre-multiplied* in alpha compositing, where their utility is now seen to be that we can vary the area coverage without changing the intensity, as with an opaque material (*key point 2*). Pre-multiplied colours are the natural choice for traditional compositing, where we mask one part of an image with another. There is no change in intensity, only in the area of colour contributing, so we need a reduced alpha.

Now suppose we place (C) at $P2 = (C, \lambda)$ or $P2' = (C, \lambda')$. This puts the original energy (C) in a smaller or larger area, according to λ. Figure 1 shows this as a vertical translation. This vertical line (C, λ) with λ varying is a line of constant energy but varying area and hence varying intensity, with lower intensity at larger λ. These are *non pre-multiplied* colours (*key point 2*). We can think of each of these as being on a vertical projection and generated from:

$$(C, 1)\begin{pmatrix} 1 & 0 \\ 0 & \lambda \end{pmatrix} = (C, \lambda)$$

Non pre-multiplied colours distribute a fixed amount of colour in varying area. If alpha increases, this effectively dilutes the colour with black (which an artist would call a "shade") and reduces the intensity of any light scattered.

To calculate which colour would produce the same intensity as $P2$ but in unit area, we use a central projection to $\alpha = 1$, as indicated in Figure 1. The result is $P3 = (C/\lambda, 1)$. This is *normalisation* in alpha compositing, here seen to be changing to a unit area at constant intensity (*key point 2*). As shown for $P2$, $\lambda < 1$ so this colour $P3$ has a higher energy than $P1 = (C, 1)$ but in the same area, resulting in a higher intensity. It will however have the same intensity as $P2$ because it is on the same radial line. We can think of the downward vertical displacement as squeezing a fixed amount of colour energy into a smaller area, giving a higher intensity. Normalisation can now be interpreted as telling us that this is the same as increasing the energy within the same area. If we start from $P2'$ where $\lambda > 1$, we get a $P3'$ with lower energy than $P1$ and corresponding comments apply.

Finally, if we place the same colour (C) at $(\lambda C, 1)$, then we have changed the amount of energy directly, while keeping the area and underlying colour unchanged. Thus $(\lambda C, 1)$ is a line of constant area but varying energy and hence intensity. These are also non pre-multiplied colours. We can think of these as being on a horizontal projection, generated from:

$$(C, 1)\begin{pmatrix} \lambda & 0 \\ 0 & 1 \end{pmatrix} = (\lambda C, 1)$$

The vertical projection is one of reducing (or increasing) the energy density by moving to a larger (or smaller) area. The horizontal projection achieves the same thing by changing the energy. In either case, the central projection (i.e. normalisation) returns that new density but in unit area. We have to exclude $\alpha = 0$ in all cases because here the area collapses and it makes no sense to talk of density. The value $\alpha = 0$ is still useful though, as we discuss in the next section.

3.1. Vanishing points, vanishing directions, ideal points

The hyperplane $(x,y,z,0)$ does not have finite points, because the scale measure w is zero and such points could not be separated. Its points do have direction however. We use this for ideal (infinite) points, which have direction but no position. We cannot use any other value of w because all such points are finite. Where two Euclidean lines intersect at (x,y,z) the homogeneous direction of this point is $(x,y,z,1)$. Parallel lines in the Euclidean plane do not intersect but they do have a definite direction, (x,y,z) say. This direction is $(x,y,z,0)$, the ideal point in the direction of the parallel lines. It is distinguished it from finite points, where w is always non-zero.

The projective space offers a particular interpretation of vanishing points. A vanishing point in a Euclidean space is the image of a point at infinity – an ideal point – in the corresponding projective space. If evenly-spaced points project closer and closer together, they are "in perspective", with the degree of perspective controllable in each dimension. The points converge on the vanishing point and the direction of projection is the vanishing direction.

Parallel lines exist in colour space. In RGB for example, the colours $(\lambda,0,0)$ and $(\lambda,1,0)$ are separated by a constant distance $(0,1,0)$. If the value of λ grows without limit, the colour coordinates of both will tend to infinity. For any finite value, they are still separated by $(0,1,0)$. Once we "reach" infinity however, the two colours become indistinguishable because they differ by a finite amount in an infinite amount. This is an ideal point and it will be in the direction of the lines. In this example the ideal point is $(1,0,0,0)$ in RGBA, with $\alpha = 0$ indicating it is ideal and $(1,0,0)$ is the vanishing direction in the Euclidean sense.

Our energy argument can be used to arrive at the same result, with reference to Figure 1. Suppose we start from $P1$, in unit area, and steadily increase its energy density (i.e. intensity). If we do this in constant area, we move horizontally to $P3$. If we do this at constant energy, we move vertically to $P2$. $P2$ and $P3$ have the same intensity because they are on the same radial line. As we tend towards infinite intensity, $P3$ tends to $(\infty,1)$ and $P2$ tends to $(C,0)$. As we do not want to compute with infinite energies, we prefer the zero area representation $(C,0)$, achieved as $P2$ goes to the limit.

A vanishing point represents a region where the unit steps of (in this case) colour measure have got closer and closer together and are ultimately no longer distinguishable. The α area varies with the colour channel. This is analogous to the w scale varying with distance. Just as we may get w wrap-around [Her92], so we may get α wrap-around. We are free to generate a vanishing direction in each dimension of our colour space, equivalent to one-, two- or three-point perspectives in 3D geometry. A practical advantage of vanishing points is that a potentially infinite colour energy range can be projected to a finite range of our choosing, perhaps to accommodate the gamut of a display.

3.2. Further properties

Normalisation in the homogeneous form is not possible when $w = 0$ because $(X,0)$ is an ideal point, a point at infinite distance in zero scale. We do however know its direction is (X). In our colour version, normalisation is not possible when $\alpha = 0$. This is because $(C,0)$ is an ideal point, a point at infinite energy density in zero area. We do however know its finite colour energy is (C) and so we know the direction. A zero-area colour contributes nothing to the final picture, so we never need to normalise it. In compositing, such pixels are discarded as "clear" image.

In our model, the energy density is unrestricted and so we can model in alpha form *all* light energies, finite and infinite, positive and negative (*key point 3*) and not just pixel values, which are what Porter and Duff model (*key point 1*). (For example, the rendering of 3D surfaces or texture mapping with varying transparent texture can now use the alpha form.) This remains true even if the areas are microscopic, to the point of being differentials.

4. Colour Calculation with PACs

With the new interpretation, we can move alpha colours beyond compositing. In this section, we show some of the possibilities. In section 5, we will also show how PACs extend what can be done with compositing. We first discuss their application to familiar general colour calculations.

4.1. Materials and illumination

It helps to think about the algebraic manipulations available with projective alpha colours. To be practical, we need these manipulations to accord with physical reality. We first think about how incoming illumination is reflected diffusely by materials to produce an outgoing colour.

We have been thinking of colour as a vector. We will now treat materials as matrices and the resultant output light as the product of the input light and an appropriate matrix. In our examples, we will use the RGBA representation for colour to make clear what the matrix can offer, but emphasise that the results do not depend on this representation. Consider the effect of the general material represented in the following matrix, when that material is illuminated by the colour vector shown.

$$(r_{in}, g_{in}, b_{in}, \alpha_{in}) \begin{pmatrix} m_{1,1}, & m_{1,2}, & m_{1,3}, & v_1 \\ m_{2,1}, & m_{2,2}, & m_{2,3}, & v_2 \\ m_{3,1}, & m_{3,2}, & m_{3,3}, & v_3 \\ t_{red}, & t_{green}, & t_{blue}, & \alpha_m \end{pmatrix}$$

This is the general projective transformation (provided in practice that it has an inverse). Graphics chip-sets already use a projective transformation for geometry; our approach means that they can use it for colour. We therefore now treat materials as a *projective transformation* over the projective alpha colours. In Section 3 the matrices used to project colours centrally, vertically and horizontally can now be thought of as materials with specialised properties. Typical materials will be diagonal matrices. The translation elements t will be zero and the righthand column will be a unit vector, giving only an affine transformation. The output red, for example, depends only on the input red and the red component of the material. Our formulation allows unbounded values.

Projective transformations of colour have not been addressed before. When one of the output channels picks up additional energy from other than its corresponding input channel, we get fluorescence. We can incorporate a colour shift operation by using the t elements in the bottom row. We can generate overall scaling of the result, such as we might need for further colour calculation, by the choice of α_m. We can also impose a vanishing direction on any or all colour channels, by the choice of v_i. v_i^{-1} is the vanishing point. The effect of two successive materials on light is achieved by multiplying their respective matrices. The order of multiplication matters in general but not for the "typical" materials mentioned in the previous paragraph.

We now consider these basic practical examples. We will revert to the single C value for colour but it will in practice expand to several channels.

4.2. Colour shift

We make a colour shift by translation:

$$(\alpha C, \alpha) \begin{pmatrix} 1 & 0 \\ t & 1 \end{pmatrix}$$

The pre-multiplied form is the colour placed at the correct scale. The alpha value of the original illumination thus scales the translation value so that it is correct for the colour space of the illumination (and therefore of the result).

Some practitioners advocate a 5×5 matrix operator for RGBA colour shift. They take the usual 4×4 matrix for RGBA operations but then extend it to include a fifth row of colour translation values. This is misguided, in the sense that we now see that the 4×4 is already in homogeneous form (*key point 4*). Rather than having the translation and the alpha separately, these terms must be multiplied in the way just described, as the homogeneous space requires.

4.3. Fluorescence

We can get cross-frequency effects by putting non-zero elements off diagonal in the colour sub-matrix (the elements $m_{i,j}$). In RGB, for example, the resultant blue can depend on the incoming red, green and blue, giving fluorescence. If energy is to be conserved, the matrix elements must be chosen accordingly but there is nothing in the mathematics to require that. It is therefore possible to invent effects which depend in imaginative ways on non-realisable materials. For example, we can construct filters which do not have a direct physical correspondence, such as one to generate a grey-scale image from a colour one (or indeed one to average the reflected colour), as Oddy and Willis did. Another example is to create fluorescence from an up-shift in frequency, from red to blue, which is the reverse of what happens in reality. The new approach, with its energy interpretation, gives us a clear basis for interpreting such operations either as physically-realisable or as for visual effect. Finally we note that matrices representing fluorescence do not commute. If M_A is a matrix which converts blue to green, and M_B converts green to red, then M_A followed by M_B converts blue to red and M_B followed by M_A returns black.

4.4. Diffuse reflection, scaling and filtering

If we wish diffusely to reflect illumination of colour $(\alpha C, \alpha)$ from a material of colour $(\alpha_m C_m, \alpha_m)$, we use:

$$(\alpha C, \alpha) \begin{pmatrix} \alpha_m C_m & 0 \\ 0 & \alpha_m \end{pmatrix} = (\alpha C \alpha_m C_m, \alpha \alpha_m)$$

The multiplication of the alphas ensures that the result is in the correct pre-multiplied form. Here C_m is "the colour" of the material, in the conventional sense, arranged as a diagonal sub-matrix rather than a vector. It consists of the elements $m_{i,i}$ of the full projective matrix and is zero elsewhere.

In our formulation, it is possible that we might have a result $\alpha > 1.0$. When we normalise this will give us exactly the same colour as we would have had without using alpha at all, so all is well. Indeed the same is true for all non-zero alpha values, including negative alpha (*key points 5,8*). For common materials, we expect the reflection coefficients C_m to be fractions but they are not required to be. Coefficients greater than one amplify the energy in that channel, as an image intensifier would.

As described, the colour which results will automatically be in the correct pre-multiplied form. If we instead adjust the bottom right α_m, without also changing the C_m multipliers, we get an overall change of scale. The material is either "super-dense", reflecting more than the incoming energy, or "super-attenuated", diluting the energy excessively. Such materials may well be more desirable than physical reality in special effects, cartoon animation and other artistic applications. We are changing the resultant area and hence the energy density of the outcome.

We can now connect our energy interpretation to the Oddy and Willis [OW91] particle model and the Wallace [Wal81] tone model. Both of these assume there are microscopic scattering centres distributed at some density across the area. This physical interpretation means that the density must be at most 1.0 and at least 0.0. Both these models assume that the particles are subsequently illuminated; that is, that the particles themselves are dark, physical material with no self-illumination. Under these circumstances we are bound to have a resultant reflection in $[0.0, 1.0]$. Porter and Duff's results are similarly based on a physical analogy, that of partial coverage by self-luminous material in a pixel-like area. The outcome is the same as for the other two models.

Calculating a filtered colour is mathematically the same as calculating one from reflected illumination. We can adjust the colour of an illuminated surface by viewing it through a filter or we can illuminate a surface with a filtered colour. Either way, the calculation is the same. The only difference lies in the way we imagine the consequences, which in turn depends on the physical arrangement being modelled. We will return to this when we consider filtered and illuminated compositing.

4.5. Multi-channel spectral colour

All the processes can be applied in the multi-channel spectral case, where there may be many channels of colour (*key point 6*). For example, a recent implementation of spectral volume rendering used 31 channels [ARC05]. With our approach, only one alpha is needed, no matter how many channels are used to represent the colour. This is not the case in current practice, as we will explain in Section 5.2.

We may include off-diagonal terms in the material matrix. Combining this with a multi-spectral representation allows such effects as showing in visible colours an image illuminated in the infra-red, as night vision binoculars would. Another example is the conversion of multi-spectral satellite images to false colour RGB.

5. Compositing with PACs

We now return to the roots of alpha, in compositing. We will show how the industry-standard 4-channel compositor can first be explained by our energy-area model and can then be extended to give filter and lighting effects, as a consequence of the multiplication present in the transformation. This will naturally offer a 16-channel compositor, needed if a full projective transformation is used but not required otherwise. We will also use the energy-area argument to show how all three earlier compositing models are variations within the theory presented here. Finally we show its use for subtractive colour, in printing.

5.1. Alpha compositing

We first rework the Porter and Duff formulation in our new way, then extend it. Conventionally, compositing is weighted addition; the usual Porter and Duff "over" operation is:

$$C_f = \alpha_A C_A + \alpha_B (1-\alpha_A) C_B$$
$$\alpha_f = \alpha_A + \alpha_B (1-\alpha_A)$$

where the subscript f indicates the light moving forward to the viewer. To get the desired "over" effect, in which the front colour partly blocks the rear, we have to reduce the area of the front colour. In projective terms, we take colour $(C_A, 1)$ and project it to $(\alpha_A C_A, \alpha_A)$. We have reduced the amount of energy which this colour contributes by restricting the area but at constant intensity. We then take colour C_B and similarly reduce its area by $(1-\alpha_A)\alpha_B$. In other words, its energy contribution is limited both by the area of "clear" area in the front layer and by its own partial "clear" area. If we wish to add these two contributions, we do it component by component. Here is the "A over B" operation, summarised in our terms.

1. A has energy C_A over unit area; that is, $(C_A, 1)$.
2. We are constructing a unit area of composite, of which A contributes a fraction α_A; so we follow the central projection in order to establish the energy of A in this reduced area: $(\alpha_A C_A, \alpha_A)$.

This is the "(energy contribution, area contribution)" for A. Now we look at what is left uncovered.

1. We have energy $(1-\alpha_A)C_B$ in area $(1-\alpha_A)$; that is, $((1-\alpha_A)C_B, (1-\alpha_A))$.
2. We need this at a fraction α_B of the area; that is we need $(\alpha_B(1-\alpha_A)C_B, \alpha_B(1-\alpha_A))$.

This is the "(energy contribution, area contribution)" for B. Hence the total contribution is, from term-by-term addition:

$$(C_f, \alpha_f) = (\alpha_A C_A + \alpha_B(1-\alpha_A)C_B, \alpha_A + \alpha_B(1-\alpha_A))$$

which is the Porter and Duff result. Incidentally, the "over" composition formula shows a natural use for negative alpha values, in scaling the amount by which we need to reduce the energy. It should also be clear that we can composite with negative colour. For example, if C_B is negative it too is reducing rather than increasing the energy contribution.

5.2. Filtered and illuminated compositing

Neither Wallace nor Porter and Duff consider filtering. Their choice of composition operators excludes this, leading to drawbacks when it is used to approximate transparency [BG04]. In fact Porter and Duff list only 12 operations, based on colour selection. Operations where both the front and back layers should simultaneously be involved are absent. In our model we can treat these as filtering operations.

The traditional "over" operation can be thought of as

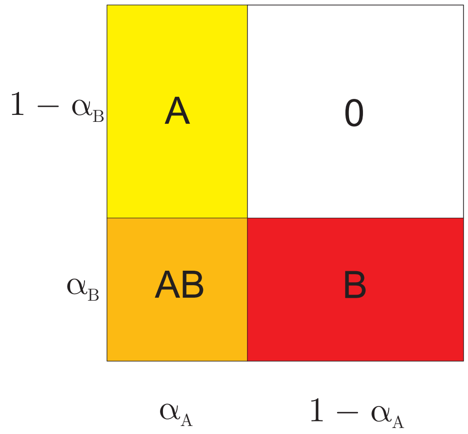

Figure 2: *Combining colours*

spray painting the rear image with the colours from the front image, with the delivered spray density limited to the alpha of the front image. Filtering cannot occur. Oddy and Willis include filtering effects, at a cost of one colour value for the filter and one for the opaque effect. This requires a 7 channel model (two RGB colours plus one value, which they call beta, to give the opaque proportion). Alternative approaches use 6 channels, with RGB and separate alphas for each. This does not conform to the 4-channel RGBA model with which it is otherwise used and so must be handled as a special case. It permits the user to set the colour to anything they choose: there is no formal basis, in contrast to our approach (*key point 7*). With multi-channel spectral images [ARC05] it is also not practicable to use one alpha per channel. For example 31 colour channels would require 31 alpha values, doubling memory and processing time. Movie post-production companies manage terabytes of data and large increases in storage and processing are commercially infeasible. With our approach, this is not needed.

Porter and Duff offer no illumination model. Wallace has a single light at the front, to fade to black at the end of a scene. Only Oddy and Willis offer a layer-by-layer illumination model, with creative possibilities. They have a pair of lights, one in front and one behind, for each layer making up the composite. In their beta model, a filter placed over an illuminated layer will change the contribution of the colour of that layer, even though the filter is not itself an illuminated element. A back-lit filter does contribute directly.

Filtering can be accomplished with layer-to-layer colour interaction through multiplication. There is only a layer-to-layer *alpha* interaction in the Porter and Duff "over" and similar operators, so we need something like the Oddy and Willis layer-to-layer *colour* interaction as well.

Figure 2 shows two alpha colours, represented with the two axes showing the alphas. This shows how the energy

	Region AB	**Region A**	**Region B**
Area:	$\alpha_A\alpha_B$	$\alpha_A(1-\alpha_B)$	$\alpha_B(1-\alpha_A)$
A over B	C_A	C_A	C_B

Table 1: *Energies of the "over" operation*

densities associated with those two alpha colours are combined in all possible ways, with no assumed ordering. We now summarise the Porter and Duff derivation using this. Then we show new filter and illumination operators in the same form, emphasising what they add to the earlier papers.

The four regions are named, following Porter and Duff, as AB, A, B, and 0. Region A is covered only by colour A, region B only by colour B, region AB by both and region 0 by neither (so we omit the latter for now). In our approach, the two colours potentially combine in the overlap region AB, for example by A filtering B. We will use "over" for illustrative purposes. With Figure 2 in mind, we can now tabulate the outcomes in each region, assuming a unit total area for simplicity (Table 1). We can readily extend this Table in the obvious way for any other operator, no matter how the colours are to be combined. To find the compositing formula for the resulting colour, we multiply each area by its colour energy, summing the total. This same Figure and Table can also be used to derive the alpha of the result. For the traditional "over", we have:

$$C_f = \alpha_A\alpha_B C_A + \alpha_A(1-\alpha_B)C_A + \alpha_B(1-\alpha_A)C_B.$$

Simplifying and similarly deriving alpha gives

$$C_f = \alpha_A C_A + \alpha_B(1-\alpha_A)C_B$$
$$\alpha_f = \alpha_A + \alpha_B(1-\alpha_A)$$

It is easy to see that α_f is all of the non-clear part of the composite and these regions are readily identified from Figure 2. This result is Porter and Duff's alpha composition. Similar results can be derived for all 12 of the operators that they consider. If we now permit filtering within *AB*, we have four cases, depending on the order of the filter (F) and the opaque layer (O).

1. When we place a filter colour *A* over a filter colour *B*, the result will be a filter and will have the (transparent) colour which is the product of the two colours.
2. When we place a filter colour *A* over an opaque colour *B*, the result will be opaque and will have the colour which is the product of the two colours.
3. When we place an opaque colour *A* over a filter colour *B*, the result will be opaque and will have the colour of *A*.
4. When we place an opaque colour *A* over an opaque colour *B*, the result will be opaque and will have the colour of *A*.

Table 2 summarises this for Regions 1-3. The underscored colours are filter colours, the rest are opaque.

Calculating a filtered colour is mathematically the same

	Region AB	**Region A**	**Region B**
Area:	$\alpha_A\alpha_B$	$\alpha_A(1-\alpha_B)$	$\alpha_B(1-\alpha_A)$
F filter F	$\underline{C_AC_B}$	$\underline{C_A}$	$\underline{C_B}$
F filter O	$\underline{C_AC_B}$	$\underline{C_A}$	C_B
O filter F	C_A	C_A	$\underline{C_B}$
O filter O	C_A	C_A	C_B

Table 2: *Energies of the "filter" operation*

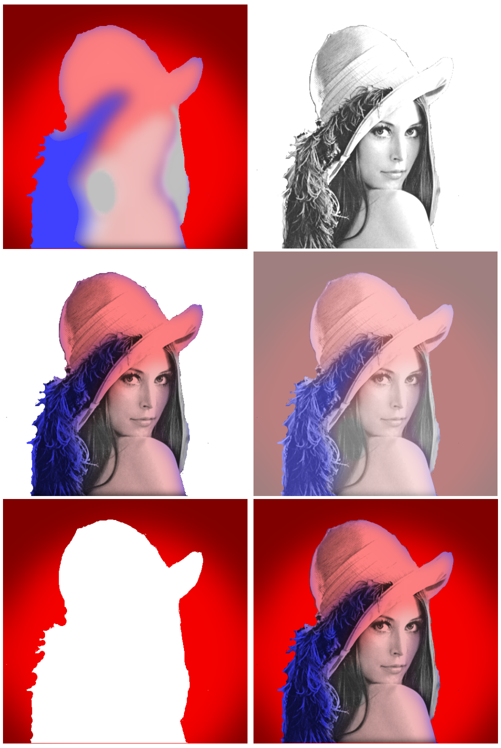

Figure 3: ***Row 1:*** *Two images A, B.* ***Row 2:*** *A filter B (c_f buffer); 50% A over B;* ***Row 3:*** *A filter B (c_r buffer); c_f over c_r.*

as calculating a reflected one. We can adjust the colour of an illuminated surface by viewing it through a filter or we can illuminate a surface with a filtered colour. However, illumination offers light energy continuing away from the viewer, potentially illuminating further layers.

To keep the algebra tidy, we will use pre-multiplied form, where $c = \alpha C$. For any operation, we will need two area-weighted colour sums, c_f for the opaque elements and c_r for the filter elements. (We use subscript r to indicate that it can affect rearward layers, further from the viewer. As before, f indicates light going forward to the viewer.) These general colour formulae depend on which O/F options we use.

1. If we choose A as a filter and B as a filter, we get
$$\begin{aligned} c_r &= c_Ac_B + c_A(1-\alpha_B) + c_B(1-\alpha_A) \\ c_f &= 0 \\ \alpha_f &= 0 \end{aligned}$$
2. If we choose A as a filter and B as opaque, we get
$$\begin{aligned} c_r &= c_A(1-\alpha_B) \\ c_f &= c_Ac_B + c_B(1-\alpha_A) \\ \alpha_f &= \alpha_B \end{aligned}$$
3. If we choose A as a opaque and B as a filter, we get
$$\begin{aligned} c_r &= c_B(1-\alpha_A) \\ c_f &= c_A \\ \alpha_f &= \alpha_A \end{aligned}$$
4. If we choose both layers to be opaque, then we get
$$\begin{aligned} c_r &= 0 \\ c_f &= c_A + c_B(1-\alpha_A) \\ \alpha_f &= \alpha_A + \alpha_B(1-\alpha_A) \end{aligned}$$

In each case we have simplified the formulae where possible. The contributions from the three non-clear Regions are distributed according to the specific case. The rearward energy can be though of as illumination which reflects from any further layers it may reach, or as a filter to any forward energy coming from further layers.

We can no longer ignore Region 0, which is clear. It will contribute to the filtering effect, by diluting the filter colour energy. It has no effect on the opaque colour, so α_f is that of the total opaque contribution and $\alpha_r = 1-\alpha_f$ in all cases. The first case gives a pure filter result. The first two cases introduce a colour change, corresponding to the effect of the front filter/light on the rear colour. This is a feature of our new approach because it derives from the colour multiplication. The second case shows colour multiplication: in the fully projective case this is matrix multiplication, with the traditional colour vector interpretation as a special case. The second and third introduce a potential to change the colour of any further layers which may later be placed behind. Again this is multiplicative. The final case gives a purely opaque result, which is the traditional "over" formula. Comparing this c_f with that in the second case, filter over opaque, we clearly see the new multiplicative term extending the Porter and Duff model. If we compare the last c_r with that of the other three, we can see that it alone does not filter. Figure 3 shows the difference between a traditional 50% alpha "over" and our new filtered/illuminated equivalent. (The monochrome face has $\alpha = 0$ in the background area.) The "over" version appears foggy and its background shows part of the red of the foreground element. In our filter/illumination version, the face has picked up the colour from the filter/illumination, with no fogginess, and the background remains clear. The third row shows the c_r buffer and the result of overlaying that on the coloured face c_f. The white area of the c_r buffer is clear, with $\alpha = 0$, due to shadowing by the face.

Each image can be an alpha colour image (C is a vector) or a projective material image (C is a matrix). These aspects are significant in extending composition. We can also control the illumination everywhere on an image with another image, not just with an overall setting for the layer. This is useful for grading an image, or to add subtle or strong lighting changes and for local tone control, as we have illustrated.

5.3. Interpretation of PAC compositing

We can composite PAC image filters and lights together to an arbitrary degree and produce a net illuminant or filter (not possible with Porter and Duff's method) for later use. This is of tremendous practical advantage because it means that we can vary the density and colour of the lighting across any area of the image, by appropriate choice of filter/light. When we subsequently composite such a filter or light onto an opaque element, we get a contribution to the final image. A filter/light could also be defined as a PAC function of space, rather than as an array of PAC pixels, or derived from a 3D rendering.

In each formula, the colour is multiplied by its alpha. This has the consequence that if we negate both the colour and its alpha, the compositing formula will be unaffected. This is consistent with the projective nature of the space.

The traditional 4-channel alpha ignores the filter/illumination energy moving to the rear. Oddy and Willis recognised the importance of the rearward-moving energy. They included both an opaque colour (the particles) and a filter colour (the medium), with a value β giving the proportion of particles; essentially (C_1, β, C_2). Our PAC interpretation explains this 7-channel beta model (*key point 7*), by making clear that we need two colours and an alpha. However, we now see that it is not essential to have this at every layer but only for every intermediate composite, of which there may only be one to render a complete stack of images. We only need a forward alpha but it makes sense to hold a rearward alpha too: practical compositors can then use either image freely. This means that all our PAC images can be in traditional 4-channel (C, α) form. At any intermediate stage, we need one image for the forward energy and one for the rearward energy. The former is the evolving image; the latter is the evolving illumination or filter.

5.4. Subtractive colours

The Euclidean plane at $\alpha = 1$ implies that the α dimension is treated differently to the colour dimensions. One outcome of this is that we do not have to explain negative alpha values physically: the projection means that any such point is the same as a corresponding positive one. It also follows that $(-C, \alpha) = (C, -\alpha)$ because both of these points are on the same central projection: Property 3 again. Similarly $(C, \alpha) = (-C, -\alpha)$. So our formulation already gives a mathematical basis for negative colour energy and negative α.

We can however provide a physical explanation (*key point 8*). In subtractive colour space, as used for printing, cyan is the ink which removes red light. Cyan might be thought of as "minus red", with magenta being "minus green" and yellow being "minus blue". If we composite with negative C but positive alpha, we can obtain a result which tells us how much ink to lay down to get that effect. As we are now subtracting from white, a result value of $\alpha = 1$ means that we have maximum ink coverage and so maximum absorption. (In contrast, we are traditionally adding to black and $\alpha = 1$ means that we have maximum screen illumination.) A lesser result means we have partial coverage, with the paper showing through the clear area and contributing to the image. Inks are either present or not, so their energies are either 1 or 0. Their areas can still be fractional, corresponding to a tint.

An example is easier to understand if we show it in RGBA form. We will assume that the background paper is maximum white over each unit area: $(1,1,1,1)$. We will consider covering a proportion α of this with cyan ink. Cyan is $(0,1,1)$. Here is the calculation, first in the familiar additive form, assuming the ink contributes α.

1. The ink has energy $(0,1,1,1)$ over unit area.
2. We need this at a fraction α of this area, so we follow the central projection in order to keep the intensity constant while varying the area: $(0, \alpha, \alpha, \alpha)$.

This is the "(energy contribution, area contribution)" for the ink. Now we look at what is left uncovered.

1. The paper has energy $(1,1,1,1)$ over unit area.
2. We need this at a fraction $(1-\alpha)$, giving a partial contribution $((1-\alpha),(1-\alpha),(1-\alpha),(1-\alpha))$.

This is the "(energy contribution, area contribution)" for the paper. The total contribution is, from term-by-term addition:

$$(C_f, \alpha_f) = ((1-\alpha), 1, 1, 1)$$

which is the resulting tone of cyan which will be printed.

We now rework this in subtractive space. Cyan ink is "minus red" so the ink now has negative red energy $(-1,0,0)$. The area will still be positive however.

1. The ink has energy $(-1,0,0,1)$ over unit area.
2. We need this at a fraction α of this area, giving $(-\alpha, 0, 0, \alpha)$.
3. The paper has energy $(1,1,1,1-\alpha)$ over the remaining area; this is our only *source* of energy, so we reduce the area but not the energy.

Adding these two contributions gives the same result as before. It is useful to compare Step 3 of the subtractive case with Step 2 for the paper in the additive case. In the additive case, we are effectively building up the unit area from two positive fractional contributions, the white of the paper and the cyan of the ink. In the subtractive case, only the paper yields positive energy, so it is present with reduced area but not with reduced energy. The ink then contributes a negative energy; that is, it reduces the total energy directly, rather than through a reduced area. Addition works by adding two lesser amounts to get a bigger total; subtraction works by starting with a bigger amount and reducing it.

By the nature of projective space, we could instead use a negative area with a positive colour for subtractive colour; or use a negative area with negative colour for additive colour. The reader may verify that these produce the result negated. A negated tuple is the same as a non-negated one because the space is projective, so the outcome is the same. All PAC calculations can be treated this way but subtractive colours offer a direct visualisation of negative colour.

6. Conclusions

Our use of a projective space and homogeneous coordinates with an energy interpretation provides a unified interpretation of a range of lighting calculations. These include diffuse reflection, filtering, fluorescence and projective materials. Other colour calculations, such as those for blue screen [SB96], can be re-interpreted in our form. Our approach removes limits on the colour or alpha values, allowing it to be used for a wide range of colour operations, not just for pixel operations in compositing. It relates pixel values to actual physical values in a unified, principled way [Bli05]. It offers an algebra, putting meaning to negative values of colour and alpha in a projective space. It explains why some *ad hoc* operations in common use are not needed and how the same results can be achieved more simply. It makes no assumptions about colour representation beyond the energy basis and so applies equally to RGB and to multi-channel spectral representation. The matrix computations are exactly those already used geometrically, the projective transformations, so a graphics card can use the same hardware for both geometric and colour calculations.

Within compositing, traditional alpha is an over-paint model, a purely additive model. We show that incorporating multiplication extends the range of possibilities while remaining within the new energy density framework. The Wallace approach is a special case of alpha and of ours. Like beta, we offer true transparency and filtering, but with only four channels. We have explained all previous compositing models in our new approach and have extended them in a natural way. We have opened up new operations from the use of a clear physical interpretation. Any non-physical operations are easy to identify as such. We can also reconcile Wallace's "tone" representation of colour, Porter and Duff's "coverage" representation and Oddy and Willis's "particle" representation: these are all different ways of saying how much incoming energy is reflected to the viewer (*key point 9*). We also include rearward energy.

The work described here amounts to a unified theory of alpha colour representation, embedded in a projective space with a clear energy interpretation, applicable to many computer graphics colour calculations.

Acknowledgements

I owe a special debt of thanks to my friend and colleague Dr John Patterson and to his home Department of Computer Science at the University of Glasgow, UK. John encouraged me to return to the area of compositing and has provided welcome feedback as the paper developed. The Department invited me for a week's stay and this was an exceptional chance to concentrate on a number of research issues which had been in the air for some time, including compositing.

I also owe many thanks to Dr Ivan Herman of the Web Consortium, for his valuable feedback on an advanced draft of the paper and for encouraging me to face some important issues more squarely. The anonymous Eurographics guide who helped this paper to its final form was also tremendously supportive and I am grateful. The final version is solely the author's responsibility.

Dedicated to the memory of Dick Grimsdale.

References

[ARC05] ABDUL-RAHMAN A., CHEN M.: Spectral volume rendering based on the Kubelka-Munk theory. *Computer Graphics Forum 24*, 3 (Sep 2005), 113–140. Proc. Eurographics Conference.

[BG04] BAUDISH P., GUTWIN C.: Multiblending: displaying overlapping windows simultaneously without the drawbacks of alpha blending. *Proc. ACM Conf. on Human Factors in Computing Systems* (April 2004), 367–374.

[Bli94] BLINN J.: Compositing, part 1: Theory. *IEEE Computer Graphics and Applications* (Sep 1994), 83–87. (Jim Blinn's Corner).

[Bli05] BLINN J.: What is a pixel? *IEEE Computer Graphics and Applications* (Sep/Oct 2005), 82–87. (Jim Blinn's Corner: page 86 in particular).

[Her92] HERMAN I.: The use of projective geometry in computer graphics. *Springer Lecture Notes in Computer Science*, 564 (1992).

[OW91] ODDY R. J., WILLIS P. J.: A physically based colour model. *Computer Graphics Forum 10*, 2 (1991), 121–127.

[PD84] PORTER T., DUFF T.: Compositing digital images. *ACM Computer Graphics 18*, 3 (Jul 1984), 253–259. (Proc. ACM SIGGRAPH 84).

[SB96] SMITH A. R., BLINN J. F.: Blue screen matting. *ACM Computer Graphics* (Aug 1996), 259–268. (Proc. ACM SIGGRAPH 96).

[Wal81] WALLACE B.: Merging and transformation of raster images for cartoon animation. *ACM Computer Graphics 15*, 3 (Aug 1981), 253–261. (Proc. ACM SIGGRAPH 81).